A Handout on Implicit Q-Learning

Rui-Teng Lin
Department of Computer Science
National Yang Ming Chiao Tung University
duck.cs09@nycu.edu.tw

1 Introduction

- This paper tackles the problem of Offline Reinforcement Learning.
- Offline Reinforcement Learning is a approach which learns from collected dataset and without interacting with the environment.
- The authors proposed a new algorithm called Implicit Q-Learning, proved that it has good convergence properties. Kostrikov et al. [2021]
- Implicit Q-Learning is a multi-step dynamic programming algorithm, and it avoids querying the out-of-distribution actions. With additional policy extraction step, it can recover near-optimal policies in the offline setting.
- In my opinion, they used some elegant and simple formulas to build up this algorithm. Though they induced more parameters, but the overall algorithm is still easy to implement.

2 Preliminaries

MDP $(\mathcal{S}, \mathcal{A}, p_0(s), p(s' \mid s, a), r(s, a), \gamma)$ where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $p_0(s)$ is distribution of initial states, $p(s' \mid s, a)$ is environment dynamics, $r(s, a)$ is reward function, and γ is discount factor.
Policy π_β is the dataset (behavior) policy.
Value functions

$$V_\tau(s) = \mathbb{E}^\tau_{a \sim \mu(\cdot|s)} \left[Q_\tau(s, a) \right]$$

$$Q_\tau(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[V_\tau(s') \right]$$

where $\mathbb{E}^\tau(x)$ is the τ^{th} expectile of x.
Loss functions

$$L_V(\psi) = \mathbb{E}_{(s,a) \sim D} \left[L_2^\tau (Q_{\hat{\theta}}(s, a) - V_\psi(s)) \right]$$

$$L_Q(\theta) = \mathbb{E}_{(s,a,s') \sim D} \left[(r(s, a) + \gamma V_\psi(s') - Q_\theta(s, a))^2 \right]$$

$$L_\pi(\phi) = \mathbb{E}_{(s,a) \sim D} \left[\exp(\beta(Q_{\hat{\theta}}(s, a) - V_\psi(s))) \log \pi_\phi(a \mid s) \right]$$

where D is the collected dataset, $L_2^\tau(x) = |\tau - \mathbb{1}(x < 0)| x^2$, $\tau \in (0, 1)$ is the asymmetric loss function and $\beta \in [0, \infty)$ is an inverse temperature.

3 Supporting Lemmas and Theoretical Analysis

Lemma 1. Let *X* and m_τ is its τ^{th} expectile be a real-valued random variable with a bounded support and supremum of the support is x^*. Then,

$$\lim_{\tau \to \infty} m_\tau = x^*$$

Proof. Expectiles of a random variable have the same supremum x^* and for all τ_1 and τ_2, we get $m_{\tau_1} \le m_{\tau_2}$. Thus, the limit follows from the properties of bounded monotonically non-decreasing functions. $\square$

Lemma 2. For all $s, \tau_1,$ and τ_2 such that $\tau_1 < \tau_2$ we get

$$V_{\tau_1}(s) \leq V_{\tau_2}(s).$$

Proof. Likely to policy improvement proof (Sutton and Barto [2018]). We can rewrite V_{τ_1} as

$$\begin{aligned}
V_{\tau_1} &= \mathbb{E}^{\tau_1}_{(a\sim\mu(\cdot|s))}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V_{\tau_1}(s')\right]\right] \\
&\leq \mathbb{E}^{\tau_2}_{a\sim\mu(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V_{\tau_1}(s')\right]\right] \\
&= \mathbb{E}^{\tau_2}_{a\sim\mu(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\mathbb{E}^{\tau_1}_{a'\sim\mu(\cdot|s')}\left[r(s',a') + \gamma\mathbb{E}_{s''\sim p(\cdot|s',a')}\left[V_{\tau_1}(s'')\right]\right]\right] \\
&\leq \mathbb{E}^{\tau_2}_{a\sim\mu(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\mathbb{E}^{\tau_2}_{a'\sim\mu(\cdot|s')}\left[r(s',a') + \gamma\mathbb{E}_{s''\sim p(\cdot|s',a')}\left[V_{\tau_1}(s'')\right]\right]\right] \\
&= \mathbb{E}^{\tau_2}_{a\sim\mu(\cdot|s)}\left[r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\mathbb{E}^{\tau_2}_{a'\sim\mu(\cdot|s')}\left[r(s',a') + \gamma\mathbb{E}_{s''\sim p(\cdot|s',a')}\mathbb{E}_{a''\sim\mu(\cdot|s'')}\left[r(s'',a'') + \ldots\right]\right]\right] \\
&\vdots \\
&\leq V_{\tau_2}(s)\square
\end{aligned}$$

Corollay 2.1. For any τ and s we have

$$V_\tau(s) \leq \max_{\substack{a\in\mathcal{A} \\ s.t. \pi_\beta(a|s)>0}} Q^*(s,a)$$

where $Q^*(s,a)$ is an optimal state-action value constrained to the dataset and defined as

$$Q^*(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[\max_{\substack{a'\in\mathcal{A} \\ s.t. \pi_\beta(a|s)>0}} Q^*(s',a')\right].$$

Proof. Convex combination is smaller than its maximum. $\square$

Theorem 3.

$$\lim_{\tau\to 1} V_\tau(s) = \max_{\substack{a\in\mathcal{A} \\ s.t. \pi_\beta(a|s)>0}} Q^*(s,a).$$

Proof. The proof can be obtained by combining **Lemma 1** and **Corollary 2.1**.

4 Discussions (Optional)

References

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning*. Adaptive Computation and Machine Learning series. Bradford Books, Cambridge, MA, 2 edition, November 2018.